Methoden in de psychologie: 2020-2021

PC-sessie 2: Lineaire regressie - interacties

Feedbackbundel

In deze oefenles maken we gebruik van de statistische software R via het programma RStudio.

De commando's die je kunt gebruiken, zijn terug te vinden in het document met voorkennis (**voorkennis_R**, zie Ufora), de nota's van **PC sessie 1** en de cursusnota's. Zorg dat je deze zaken goed doornomen hebt!

Om je op weg te helpen voor deze specifieke oefenles, hebben we het script **Opgave_script_PC2.R** op Ufora ter beschikking gesteld.

Oefening 1: Predictoren voor stress

In deze oefening maken we opnieuw gebruik van de data uit een studie die de invloed van een aantal variabelen op stress nagaat ($n = 377$). Zie **PC sessie 1** voor de uitleg over de data en de verschillende variabelen.

```
# path waar je databestand staat correct specificeren
# of working directory in R wijzigen
# datastress<-read.csv("DataStress.csv",sep=";",dec=".")

# Data kunnen ook als volgt ingelezen worden:
datastress<-read.csv("http://www.da.ugent.be/datasets/DataStress.csv",
                     sep=";",dec=".")
class(datastress)
```

```
[1] "data.frame"
```

```
dim(datastress)
```

```
[1] 377  11
```

```
head(datastress)
```

```
     SEX EDUCLEV CHI      AGE DECAUT JOBINS PSYDEM TSOSUP DEPRES SKILLUT
1    man    laag   2 59.45893     40    3.0     34     30     20       0
2    man    hoog   6 46.40298     48    3.0     34     28     19       0
3    man    hoog   3 38.74179     40    6.0     32     24     32      -2
4    man  middel   0 52.78679     40    4.5     25     14     18      -2
5 vrouw  middel   7 39.89203     36    4.5     24     15     25       0
6 vrouw  middel   2 36.72416     44    6.0     26     30     26       0
     STRESS
1  41.43302
2  99.78582
3  43.82572
4  16.74125
5 147.29815
6  80.11587
```

We starten met het lineair regressiemodel met stress *als uitkomstvariabele en vijf predictoren, namelijk* chi*,* age*,* depres*,* decaut *en* tsosup*.*

1. Ga na of een model met een interactie tussen age *en* decaut *de variantie in de uitkomstvariabele* stress *beter kan verklaren dan een model zonder deze interactie.*

(a) Welk lineair model schatten we om na te gaan of de interactie bijdraagt tot het model? Schrijf dit neer en interpreteer de parameters van het model.

$$\texttt{stress}_i = \beta_0 + \beta_1 \times \texttt{chi}_i + \beta_2 \times \texttt{dep}_i + \beta_3 \times \texttt{tsosup}_i + \beta_4 \times \texttt{age}_i + \beta_5 \times \texttt{decaut} + \beta_6 \times \texttt{age}_i \times \texttt{decaut}_i + \epsilon_i$$

In dit model is het effect van decaut op stress gelijk aan $\beta_5 + \beta_6 \times \texttt{age}_i$. Dit betekent dat de verandering in de verwachte waarde van stress wanneer decaut toeneemt met 1 eenheid terwijl alle overige predictoren constant blijven, afhangt van het niveau waarop leeftijd (age=$\texttt{age}_i$) constant gehouden wordt. β_5 is het hoofdeffect van decaut, dit is het effect van decaut wanneer age=0 (heeft dit effect een zinvolle praktische interpretatie?). β_6 stelt het interactie-effect voor.

Analoog kunnen we afleiden dat het effect van age gelijk is aan $\beta_4 + \beta_6 \times \texttt{decaut}_i$ en dus afhangt van het niveau waarop decaut constant gehouden wordt.

(b) Wat is de nulhypothese die hoort bij de toets die we gaan uitvoeren?

Aan de hand van bovenstaand lineair model kunnen we volgende nulhypothese formuleren: 'Er is geen interactie-effect tussen leeftijd en decision autonomy'. In symbolen uitgedrukt is dit: $H_0 : \beta_6 = 0$. De alternatieve hypothese is $H_1 : \beta_6 \neq 0$.

(c) Voer de toets uit in R. Rapporteer de geobserveerde toetsingsgrootheid en formuleer een conclusie.

Een toets voor een interactie-effect kunnen we altijd op drie manieren uitvoeren. Deze drie methodes zullen tot hetzelfde besluit leiden omtrent het getoetste interactie-effect.

Methode 1: we kijken naar het overzicht van alle toetsen van de predictoren in ons model via `Anova()`.

```
library(car)
# Voeg interactie toe in het model
ModelStress1 <- lm(STRESS ~ CHI + DEPRES + TSOSUP + AGE + DECAUT
                   + AGE:DECAUT, data = datastress)

# Toets voor interactie via Anova-commando
  # Merk op: alle predictoren zijn van intervalniveau
  # Indien nominale predictoren in model: eerst op effect-codering overgaan!
Anova(ModelStress1, type = 3)
```

```
 Anova Table (Type III tests)

Response: STRESS
             Sum Sq  Df   F value    Pr(>F)
(Intercept)    603   1    2.7845  0.096027 .
CHI         240587   1 1110.5430 < 2.2e-16 ***
DEPRES         731   1    3.3758  0.066962 .
TSOSUP        2265   1   10.4543  0.001334 **
AGE            260   1    1.1988  0.274282
DECAUT         763   1    3.5198  0.061426 .
AGE:DECAUT     879   1    4.0576  0.044696 *
Residuals    80156 370
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We stellen vast dat op het 5% significantieniveau de interactie tussen age en decaut significant verschillend is van 0 ($F(1, 370) = 4.058, p < 0.05$). We verwerpen de nulhypothese: we besluiten op het 5% significantieniveau dat een model met de interactieterm de variatie in de uitkomstvariabele stress beter verklaart dan een model zonder de interactieterm.

Methode 2: we kunnen hetzelfde resultaat bekomen aan de hand van een modelvergelijkingstoets via het commando `anova()`.

```
ModelStress2 <- lm(STRESS ~ CHI + DEPRES + TSOSUP + AGE + DECAUT,
                   data = datastress)
anova(ModelStress2, ModelStress1)
```

```
 Analysis of Variance Table

Model 1: STRESS ~ CHI + DEPRES + TSOSUP + AGE + DECAUT
Model 2: STRESS ~ CHI + DEPRES + TSOSUP + AGE + DECAUT + AGE:DECAUT
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    371 81035
```

```
2    370 80156  1    879.03 4.0576 0.0447 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ook hier stellen we vast dat op het 5% significantieniveau de interactie tussen age en decaut significant verschillend is van 0 met exact dezelfde toetsingsgrootheid die gerapporteerd wordt: $(F(1, 370) = 4.058, p < 0.05)$. We verwerpen de nulhypothese: we besluiten op het 5% significantieniveau dat een model met de interactieterm de variatie in de uitkomstvariabele stress beter verklaart dan een model zonder de interactieterm.

Methode 3: aangezien er in het geval van een interactie tussen twee continue predictoren bij de toets voor het interactie-effect slechts 1 parameter getoetst wordt (i.e. 1 vrijheidsgraad voor de teller bij de F-toets), kunnen we het resultaat ook aflezen bij de t-toets voor het interactie-effect via het commando `summary()` dat toegepast wordt op het gefitte model.

```
summary(ModelStress1)
```

```
Call:
lm(formula = STRESS ~ CHI + DEPRES + TSOSUP + AGE + DECAUT +
    AGE:DECAUT, data = datastress)

Residuals:
    Min      1Q  Median      3Q     Max
-39.451 -10.377   0.562   9.412  40.971

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.86582   25.08924   1.669  0.09603 .
CHI          9.84907    0.29555  33.325  < 2e-16 ***
DEPRES      -0.27787    0.15123  -1.837  0.06696 .
TSOSUP      -0.71690    0.22172  -3.233  0.00133 **
AGE          0.60265    0.55043   1.095  0.27428
DECAUT       1.31266    0.69967   1.876  0.06143 .
AGE:DECAUT  -0.03105    0.01542  -2.014  0.04470 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.72 on 370 degrees of freedom
Multiple R-squared:  0.8195,    Adjusted R-squared:  0.8166
F-statistic:   280 on 6 and 370 DF,  p-value: < 2.2e-16
```

We lezen het volgende resultaat af voor de t-toets: $t(370) = -2.014, p = 0.0447$. Dit komt inderdaad overeen met bovenstaande resultaten. Er geldt dat $(-2.014)^2$ op afronding na gelijk is aan de geobserveerde toetsingsgrootheid bij de F-toets. De p-waarde voor de t-toets is exact gelijk aan die voor de F-toets.

Parameterschattingen: in bovenstaande output lezen we ook de geschatte regressieparameters af. We schatten dat wanneer age met 1 jaar toeneemt, terwijl decaut=0 en alle overige predictoren constant blijven, de gemiddelde stress stijgt met 0.6 eenheden (i.e. $0.6-0.031\times 0 = 0.6$). Wanneer decaut=40 en alle overige predictoren constant blijven, schatten we dat de gemiddelde stress daalt met 0.637 eenheden (i.e. $0.6 - 0.031 \times 40 = -0.637$).
We zien dus inderdaad dat het effect van age afhankelijk is van decaut (en omgekeerd).

2. In een volgende stap staan we stil bij mogelijke problemen met multicollineariteit in het bovenstaande model.

(a) Bekijk de correlaties tussen de verschillende predictoren (zonder de interactie). Wat merk je?

In **PC sessie 1** bekeken we reeds de correlatie tussen de verschillende continue variabelen in de dataset:

```
cor(datastress[,c(3:11)])
```

```
                CHI         AGE      DECAUT      JOBINS      PSYDEM      TSOSUP
CHI      1.00000000  0.07148634 -0.18479520  0.10249388  0.16297582 -0.10903034
AGE      0.07148634  1.00000000  0.02144086 -0.07736374  0.04587973 -0.02220340
DECAUT  -0.18479520  0.02144086  1.00000000 -0.19338694  0.14434444  0.33600736
JOBINS   0.10249388 -0.07736374 -0.19338694  1.00000000  0.09922961 -0.23318572
PSYDEM   0.16297582  0.04587973  0.14434444  0.09922961  1.00000000 -0.13175262
TSOSUP  -0.10903034 -0.02220340  0.33600736 -0.23318572 -0.13175262  1.00000000
DEPRES   0.57253573 -0.01132709 -0.17030452  0.23951162  0.26458773 -0.17346256
SKILLUT  0.02991649  0.18390638  0.08750807 -0.01365915  0.10826743  0.06798142
STRESS   0.89627359 -0.02049622 -0.20506192  0.06113044  0.13431119 -0.16706880
             DEPRES     SKILLUT      STRESS
CHI      0.57253573  0.02991649  0.89627359
AGE     -0.01132709  0.18390638 -0.02049622
DECAUT  -0.17030452  0.08750807 -0.20506192
JOBINS   0.23951162 -0.01365915  0.06113044
PSYDEM   0.26458773  0.10826743  0.13431119
TSOSUP  -0.17346256  0.06798142 -0.16706880
DEPRES   1.00000000  0.03373761  0.49166640
SKILLUT  0.03373761  1.00000000  0.01843136
STRESS   0.49166640  0.01843136  1.00000000
```

Aangezien de predictoren in het model chi, age, depres, decaut en tsosup zijn, concentreren we ons hier op de samenhang tussen deze variabelen.

We zien geen extreem grote correlaties tussen de verschillende predictoren, enkel de correlatie tussen depres en chi is vrij hoog (0.57).

Merk op dat de afwezigheid van hoge correlaties tussen predictoren niet voldoende is om multicollineariteit uit te sluiten. We bekijken in een volgende vraag expliciet of er sprake is van multicollineariteit tussen predictoren.

(b) Maak zelf een variabele aan die in bovenstaand model codeert voor de interactie tussen age en decaut. Bereken de correlatie tussen deze variabele en respectievelijk de variabelen age en decaut. Wat stel je vast?

```
# variabele die codeert voor een interactie-effect is product van de 2 variabelen
int_age_decaut <- datastress$AGE * datastress$DECAUT

cor(datastress$AGE, int_age_decaut)
```

```
[1] 0.5512012
```

```
cor(datastress$DECAUT, int_age_decaut)
```

```
[1] 0.8371575
```

We observeren een hoge correlatie tussen decaut en de interactieterm (0.84). De correlatie tussen age en de interactieterm is kleiner, maar nog steeds substantieel (0.55). In een volgende vraag bekijken we expliciet of deze hoge correlaties voor problemen zorgen met multicollineariteit bij het fitten van het model.

(c) Ga na of er potentiële problemen zijn met multicollineariteit in het bovenstaande model? Wat zijn je bevindingen?

```
# Variance inflation factors (VIFs) bekijken om problemen met
# multicollineariteit in kaart te brengen
vif(ModelStress1)
```

```
      CHI     DEPRES     TSOSUP        AGE     DECAUT AGE:DECAUT
 1.544431   1.526936   1.161252  21.037571  49.223143  70.094393
```

De hoge VIFs (21.04 voor age , 49.22 voor decaut en 70.09 voor de interactieterm) wijzen erop dat de predictorvariabelen age, decaut en de interactieterm elkaar onderling sterk voorspellen. Dit zorgt voor onnauwkeurige parameterschattingen (gekenmerkt door een grote standaardfout). Als vuistregel geldt: indien de VIF onder 10 blijft, zijn er geen noemenswaardige problemen.

(d) Maak 2 nieuwe variabelen aan: age_c en decaut_c die de gecentreerde scores van respectievelijk age en decaut bevatten. Maak ook een variabele aan die codeert voor de interactie tussen age_c en decaut_c.

Als je de correlatie tussen beide gecentreerde variabelen en hun interactieterm beschouwt, wat stel je vast?

```
# centreer de variabelen age en decaut
age_c <- datastress$AGE - mean(datastress$AGE)
decaut_c <- datastress$DECAUT - mean(datastress$DECAUT)
```

```r
# maak een interactieterm aan voor de 2 gecentreerde variabelen
int_agec_decautc <- age_c*decaut_c

# bereken correlaties
cor(age_c,decaut_c)
```

```
[1] 0.02144086
```

```r
cor(age_c,int_agec_decautc)
```

```
[1] 0.03868097
```

```r
cor(decaut_c,int_agec_decautc)
```

```
[1] -0.06377139
```

We zien dat de correlatie tussen de gecentreerde variabelen voor age en decaut ongewijzigd is (0.021), maar dat de correlaties tussen de gecentreerde variabelen en de interactieterm sterk gedaald zijn. De correlatie tussen de gecentreerde decaut en de interactieterm is gedaald naar -0.06 (tegenover 0.84) en ook de correlatie tussen de gecentreerde age en de interactieterm is klein geworden (0.04 tegenover 0.55).

3. Vervang in het oorspronkelijk model met de interactie tussen age en decaut de variabelen age en decaut door de respectievelijke gecentreerde variabelen.

(a) Welk lineair model schatten we om na te gaan of de gecentreerde interactie bijdraagt tot het model? Schrijf dit neer en interpreteer de parameters van het model.

$$\begin{aligned} \texttt{stress}_i \;=\; & \beta_0 + \beta_1 \times \texttt{chi}_i + \beta_2 \times \texttt{dep}_i + \beta_3 \times \texttt{tsosup}_i + \\ & \beta_4 \times \texttt{age\_c}_i + \beta_5 \times \texttt{decaut\_c}_i + \\ & \beta_6 \times \texttt{age\_c}_i \times \texttt{decaut\_c}_i + \epsilon_i \end{aligned}$$

In dit model is het effect van decaut op stress gelijk aan $\beta_5 + \beta_6 \times \texttt{age\_c}_i$. Dit betekent dat de verandering in de verwachte waarde van stress wanneer decaut toeneemt met 1 eenheid terwijl alle overige predictoren constant blijven, afhangt van het niveau waarop de gecentreerde leeftijd (age_c=$\texttt{age\_c}_i$) constant gehouden wordt. β_5 is het hoofdeffect van decaut, dit is het effect van decaut wanneer age_c=0. De interpretatie van het hoofdeffect wordt praktisch gezien zinvoller wanneer de variabelen gecentreerd zijn, het is namelijk het effect van decaut wanneer de leeftijd gelijk is aan de gemiddelde leeftijd (geobserveerde steekproefgemiddelde van age). β_6 stelt het interactie-effect voor.

Analoog kunnen we afleiden dat het effect van age gelijk is aan $\beta_4 + \beta_6 \times \texttt{decaut\_c}_i$ en dus afhangt van het niveau waarop de gecentreerde decaut (decaut_c=$\texttt{decaut\_c}_i$) constant gehouden wordt.

(b) Voer de toets uit in R en formuleer je conclusie.

Methode 1: we kijken naar het overzicht van alle toetsen van de predictoren in ons model via `Anova`:

```
ModelStress3 <- lm(STRESS ~ CHI + DEPRES + TSOSUP + age_c + decaut_c
                   + age_c:decaut_c, data = datastress)
Anova(ModelStress3, type = 3)
```

```
 Anova Table (Type III tests)

Response: STRESS
               Sum Sq  Df   F value    Pr(>F)
(Intercept)     22831   1  105.3860 < 2.2e-16 ***
CHI            240587   1 1110.5430 < 2.2e-16 ***
DEPRES            731   1    3.3758 0.0669618 .
TSOSUP           2265   1   10.4543 0.0013336 **
age_c            3239   1   14.9504 0.0001304 ***
decaut_c          149   1    0.6896 0.4068329
age_c:decaut_c    879   1    4.0576 0.0446964 *
Residuals       80156 370
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We stellen vast dat op het 5% significantieniveau de gecentreerde interactie tussen age en decaut significant verschillend is van 0 met de toetsingsgrootheid: $(F(1, 370) = 4.058, p < 0.05)$. We verwerpen de nulhypothese: we besluiten op het 5% significantieniveau dat een model met de interactieterm de variatie in de uitkomstvariabele stress beter verklaart dan een model zonder de interactieterm.

Methode 2: we kunnen hetzelfde resultaat bekomen aan de hand van een modelvergelijkingstoets via het commando `anova()`.

```
ModelStress4 <- lm(STRESS ~ CHI + DEPRES + TSOSUP + age_c + decaut_c,
                   data = datastress)
anova(ModelStress4, ModelStress3)
```

```
 Analysis of Variance Table

Model 1: STRESS ~ CHI + DEPRES + TSOSUP + age_c + decaut_c
Model 2: STRESS ~ CHI + DEPRES + TSOSUP + age_c + decaut_c + age_c:decaut_c
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    371 81035
2    370 80156  1    879.03 4.0576 0.0447 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ook hier stellen we vast dat op het 5% significantieniveau de gecentreerde interactie tussen `age` en `decaut` significant verschillend is van 0 met exact dezelfde toetsingsgrootheid die gerapporteerd wordt: $(F(1, 370) = 4.058, p < 0.05)$.

Methode 3: we kunnen het resultaat ook aflezen bij de t-toets voor het interactie-effect via het commando `summary()` dat toegepast wordt op het gefitte model.

```
summary(ModelStress3)
```

```
Call:
lm(formula = STRESS ~ CHI + DEPRES + TSOSUP + age_c + decaut_c +
    age_c:decaut_c, data = datastress)

Residuals:
    Min      1Q  Median      3Q     Max
-39.451 -10.377   0.562   9.412  40.971

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    65.99634    6.42878  10.266  < 2e-16 ***
CHI             9.84907    0.29555  33.325  < 2e-16 ***
DEPRES         -0.27787    0.15123  -1.837  0.06696 .
TSOSUP         -0.71690    0.22172  -3.233  0.00133 **
age_c          -0.46721    0.12083  -3.867  0.00013 ***
decaut_c       -0.08927    0.10750  -0.830  0.40683
age_c:decaut_c -0.03105    0.01542  -2.014  0.04470 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.72 on 370 degrees of freedom
Multiple R-squared:  0.8195,    Adjusted R-squared:  0.8166
F-statistic:   280 on 6 and 370 DF,  p-value: < 2.2e-16
```

We lezen het volgende resultaat af voor de t-toets: $t(370) = -2.014, p = 0.0447$. Dit komt inderdaad overeen met bovenstaande resultaten.

We merken op dat de toets voor het interactie-effect ongewijzigd blijft wanneer de variabelen gecentreerd zijn.

Parameterschattingen: we stellen wel vast dat de parameterschattingen en bijhorende toetsen voor de hoofdeffecten gewijzigd zijn.

De parameterschatting voor age kunnen we als volgt interpreteren:

$$\beta_4 + \beta_6 \times \texttt{decaut\_c}_i$$

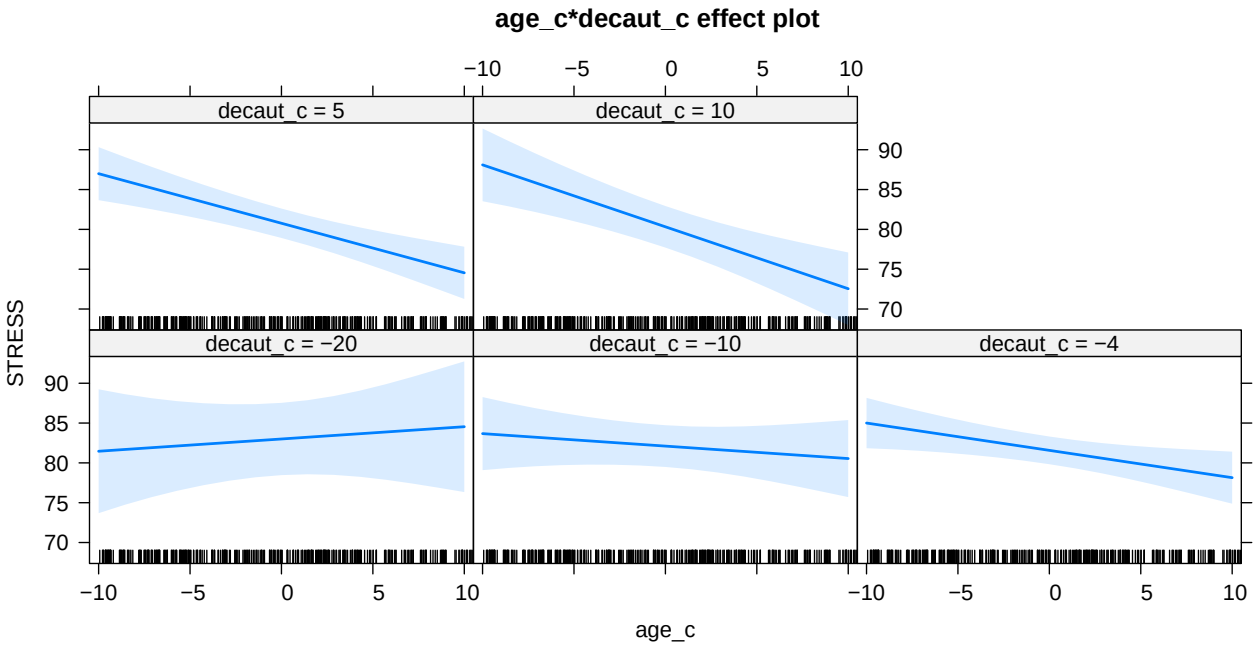
$$-0.47 + (-0.03) \times \texttt{decaut\_c}_i$$

- **Indien decaut_c=0**
  Wanneer decaut_c=0 impliceert dit dat er een gemiddeld niveau is van decision autonomy (omwille van het centreren).
  Wanneer decaut_c=0 is, wordt de bovenstaande formule $-0.47 + (-0.03) \times 0 = -0.47$.
  De bijhorende interpretatie is dan:
  We schatten dat, wanneer alle overige predictoren constant blijven en age met 1 jaar stijgt, de gemiddelde stress met 0.467 eenheden afneemt als decaut gelijk is aan het steekproefgemiddelde.
  Of op een andere manier verwoord: op basis van het model schatten we dat een stijgende leeftijd gemiddeld genomen gepaard gaat met een dalende stressscore op het gemiddelde niveau van decision autonomy (en alle overige predictoren constant).

- **Indien decaut_c≠0**
  Ter illustratie nemen we de waarde decaut_c=10, de interpretatie kan aangepast worden naar eender welk getal.
  Wanneer decaut_c=10 impliceert dit dat de waarde voor decision autonomy 10 eenheden boven het gemiddelde ligt (omwille van het centreren).
  Wanneer decaut_c=10 is, wordt de bovenstaande formule $-0.47 + (-0.03) \times 10 = -0.777$.
  De bijhorende interpretatie is dan:
  Wanneer decaut 10 eenheden boven het gemiddelde ligt (decaut_c=10), dan schatten we dat de gemiddelde stress met 0.777 eenheden daalt als age met 1 eenheid stijgt en alle overige predictoren constant blijven.

We kunnen het interactie-effect tussen age_c en decaut_c visualiseren via een effecten-plot.

```
library(effects)
plot(effect("age_c:decaut_c",ModelStress3))
```



We zien hier dus duidelijk dat het effect van age_c afhankelijk is van het niveau waarop decaut_c constant gehouden wordt: de helling van de regressierechte die het effect van age_c op stress voorstelt, verandert voor verschillende waarden voor decaut_c.

(c) Ga opnieuw na of er mogelijks problemen zijn met multicollineariteit.

```
vif(ModelStress3)
```

```
            CHI          DEPRES          TSOSUP           age_c       decaut_c
       1.544431        1.526936        1.161252        1.013850        1.161902
age_c:decaut_c
       1.034780
```

We stellen geen problemen meer vast met multicollineariteit (alle VIFs dicht bij 1).

Indien een interactieterm in een lineair regressiemodel wordt opgenomen, is het centreren of standaardiseren van continue predictoren niet strikt noodzakelijk. Het kan echter wel eventuele problemen met multicollineariteit verhelpen en het vergemakkelijkt de interpretatie van de hoofdeffecten.

Oefening 2: pijneducatie

In deze oefening maken we opnieuw gebruik van de data uit de studie rond pijneducatie $n = 121$. Zie de cursunota's en **PC sessie 1** voor de uitleg over de data en de verschillende variabelen.

```
# path waar je databestand staat correct specificeren
# of working directory in R wijzigen
# pijneducatie<-read.csv("pijneducatie.csv",sep=";",dec=".")

# Data kunnen ook als volgt ingelezen worden:
pijneducatie<-
read.csv("http://www.da.ugent.be/datasets/pijneducatie.csv",
          sep=";",dec=".")
class(pijneducatie)
```

```
[1] "data.frame"
```

```
dim(pijneducatie)
```

```
[1] 121   5
```

```
head(pijneducatie)
```

```
         Buig Gender    Leeft        Conditie       Dep
1 -10.863835    Man 39.51317 Rugpijneducatie 33.214520
2   1.699991  Vrouw 40.25680 Rugpijneducatie 23.221516
3   5.249518    Man 35.34418 Rugpijneducatie 35.309551
4   9.263604  Vrouw 33.27902 Rugpijneducatie 15.787603
5   9.523696    Man 39.63532 Rugpijneducatie  8.411790
6  31.783503  Vrouw 25.09226 Rugpijneducatie  1.546317
```

Fit een model voor de uitkomst Buig (i.e. verschilscore in voorover buigen) met een interactie tussen de nominale predictoren Conditie en Gender, waar ook gecontroleerd wordt voor Dep en Leeft. We gaan na of er een effect is van Conditie op Buig. Aangezien Conditie zowel als hoofdeffect als in een interactie met Gender in het model opgenomen is, beschouwen we toetsen voor de volgende effecten in het model: hoofdeffect van Conditie en interactie-effect tussen Conditie en Gender.
*In **Bordsessie 2** komen we ook op deze oefening terug en zullen we dieper ingaan op de interpretatie van de parameters.*

1. Schrijf het bovenstaande lineair regressiemodel neer.

Conditie bestaat uit 3 niveaus; laat $HV1$ en $HV2$ de 2 hulpveranderlijken voorstellen die coderen voor deze variabele.

$$\begin{aligned} \texttt{Buig}_i &= \beta_0 + \beta_1 \times \texttt{Dep}_i + \beta_2 \times \texttt{Leeft}_i \\ &\quad + \beta_3 \times \texttt{Gender}_i + \beta_4 \times \texttt{HV1}_i + \beta_5 \times \texttt{HV2}_i \\ &\quad + \beta_6 \times \texttt{HV1}_i \times \texttt{Gender}_i + \beta_7 \times \texttt{HV2}_i \times \texttt{Gender}_i + \epsilon_i \end{aligned}$$

2. Geef de nul- en alternatieve hypotheses die horen bij de toets voor het hoofdeffect en de toets voor het interactie-effect.

- **Hoofdeffect**
  De nulhypothese stelt dat er geen hoofdeffect is van Conditie: $H_0 : \beta_4 = \beta_5 = 0$. De alternatieve hypothese H_1 stelt dat minstens 1 van de β's verschillend is van 0.
- **Interactie-effect**
  De nulhypothese stelt dat er geen interactie-effect is tussen Conditie en Gender: $H_0 : \beta_6 = \beta_7 = 0$. De alternatieve hypothese H_1 stelt dat minstens 1 van de β's verschillend is van 0.

3. Voer de toetsen uit in R en formuleer een conclusie.

Aangezien zowel de toets voor het hoofdeffect als de toets voor het interactie-effect inhoudt dat meerdere parameters simultaan getoetst worden (doordat Conditie een nominale variabele is die bestaat uit 3 niveaus), maken we gebruik van het commando `Anova()`. Voor het toetsen zorgen we ervoor dat de nominale variabelen gecodeerd worden aan de hand van effect-codering.

```
library(car)
pijneducatie_interactie_test <- lm(Buig ~ Dep + Leeft + Gender + Conditie +
                                     Gender:Conditie, data=pijneducatie,
                                   contrasts=list(Conditie=contr.sum,Gender=contr.sum))
Anova(pijneducatie_interactie_test,type=3)
```

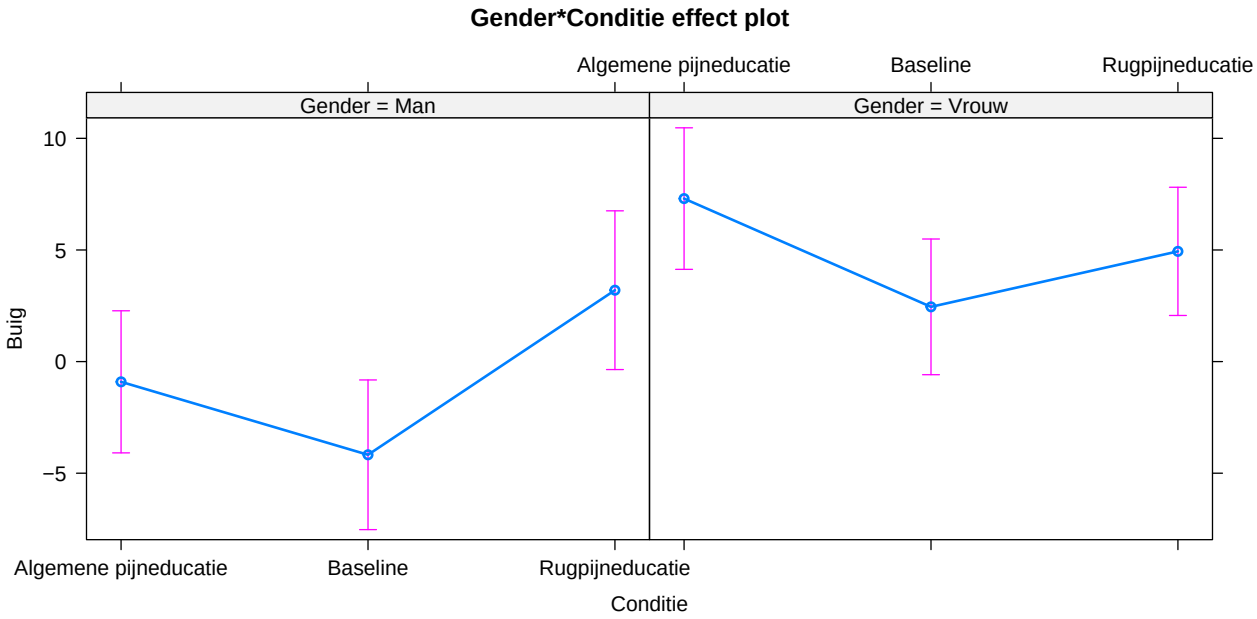
```
 Anova Table (Type III tests)

Response: Buig
               Sum Sq  Df F value    Pr(>F)
(Intercept)    1211.3   1 23.6766 3.723e-06 ***
Dep            4675.2   1 91.3818 3.189e-16 ***
Leeft            64.0   1  1.2505  0.265833
Gender          901.8   1 17.6265 5.382e-05 ***
Conditie        546.4   2  5.3404  0.006079 **
```

```
Gender:Conditie  220.5   2  2.1554  0.120597
Residuals       5781.2 113
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

De toets voor het hoofdeffect van Conditie toetst het effect van Conditie, uitgemiddeld over de niveaus van Gender. Er is een significant hoofdeffect van Conditie ($F(2, 113) = 5.34, p = 0.006$) maar geen statistisch significante interactie met Gender op het 5% significantie-niveau ($F(2, 113) = 2.155, p = 0.121$) in dit model.

4. Maak een effectenplot om het interactie-effect tussen Conditie en Gender te interpreteren.

```
library(effects)
plot(effect("Gender:Conditie",pijneducatie_interactie_test))
```



We stellen vast dat zowel bij mannen als vrouwen de gemiddelde verschilscores (Buig, conditioneel op Dep en Leeft) het laagst zijn in de baseline. Verder observeren we dat de trend voor de andere gemiddeldes gelijkaardig is voor mannen en vrouwen; dit stemt overeen met de conclusie dat er geen evidentie is voor een interactie-effect tussen Gender en Conditie. We stellen wel vast dat de geschatte gemiddelde verschilscores voor alle condities hoger zijn bij vrouwen dan bij mannen.